基于 CRFs 和词典信息的中古汉语自动分词*

王晓玉　李　斌
(南京师范大学文学院　南京 210097)

摘要:【目的】验证中古时期分词一致性和语料类别对 CRFs 分词效率的影响, 在此基础上进一步提高分词效率, 降低人工校对的工作量。【方法】以中古时期的史书、佛经、小说类语料为例, 针对中古汉语的自动分词问题, 优化分词原则, 运用 CRFs 模型和词典相结合的方法, 消除中古汉语人工分词结果中易出现的分词不一致问题; 同时在 CRFs 分词中引入字符分类、字典信息两种特征, 并通过对比实验选取每种特征最合适的分词模板。【结果】实验结果显示, 分词结果的总 F 值在封闭测试中达到 99%以上, 开放测试的综合测试中也达到 89%-95%。【局限】分词不一致研究主要针对双字词, 因此三字以上词语(多字词)的识别效果稍有欠缺。【结论】在有效提高分词一致性的前提下, 字符分类、词典标记特征能够有效提高中古汉语 CRFs 分词的精确度。同时本文提出的中古汉语分词系统可以服务于中古时期多类别的汉语语料。

关键词: CRFs 模型　分词一致性　中古汉语　自动分词
分类号: TP391

1　引　言

众所周知, 汉语中词与短语之间的界限往往难以划分清楚, 这一现象在中古汉语中更为突出。在汉语史上, 中古是指东汉末年至隋朝这段时期, 此时期汉语正处于质变期, 由于汉语中的词汇在上古时期以单字词为主, 在近代时期以双字词为主, 而中古时期汉语正处于由单字词为主向双字词为主转变的过渡期, 该过渡期中存在大量状态介于词和短语之间的字组, 这些字组的情况各不相同, 有的字组正处于词汇化的进程中, 有的则是由多个汉字临时组合起来作为词使用。正是这些字组的存在, 使得中古时期词和短语的边界更加不明确。在构建中古语料库的分词阶段, 由于各人语感不同, 再加上这些字组发生词汇化的时期、在具体文献中的词汇化程度难以完全量化, 这必然会造成中古汉语分词上的困难, 不仅直接导致了人工分词结果中出现的分词不一致现象, 在以人工分词结果为训练语料的前提下, 也严重制约着机器分词准确率和一致性的提高。而分词在汉语语料库建设中是一项基础性工作, 它对后续标注、语义分析等起着至关重要的作用。

中古时期的汉语语料相比现代来说不算多, 但其规模也在数亿字以上, 包含大量史书、佛经及民间文学、杂著类语料[1]。用计算机处理中古语料时不可避免地要进行中古汉语分词。然而, 目前与古代汉语信息处理相关的研究成果本就不多, 与中古汉语相关的分词研究更加少见。王嘉灵[2]基于《汉书》进行自动分词研究, 制定了《汉书》分词规范, 并在加入字符分类、上古音韵特征的基础上, 用 CRFs 模型进行分词实验, 实验结果的 F 值达到 94.4%, 但该研究仅以《汉书》为自动分词的实验语料, 一本书难以代表中古时期的语料全貌, 再加上中古时期语料类别有很多, 各类别语料间差异比较大, 不仅史书、佛经、杂著等类别间存在差异, 各类别内部, 如佛经语料内部的译经

通讯作者: 李斌, ORCID: 0000-0002-7328-9947, E-mail: libin.njnu@gmail.com。

*本文系国家社会科学基金重大项目“汉语史研究语料库建设研究”(项目编号: 10&ZD117)、教育部人文社会科学青年项目“汉语历时词汇数据库的构建与计量研究”(项目编号: 16YJC740034)和国家社会科学基金重大项目“基于《汉学引得丛刊》的典籍知识库构建及人文计算研究”(项目编号: 15ZDB127)的研究成果之一。

和僧传间也存在词汇上的差异，这也使得该研究成果的可扩展性大大降低。王晓玉等[3]从中古时期具有代表性的语料中抽样选取 28 万余字作为实验对象，统计这些语料人工分词结果中出现的切分错误、分词不一致、组合型歧义字串的数目及比例，着重研究分词不一致字串产生的原因和分类，并提出解决方案的设想，该研究覆盖了中古时期的佛经、史书、小说类语料，不仅呈现了中古汉语人工分词的大体概貌，并找出分词中具体存在的问题，为进一步研究奠定了基础。

在以上研究的基础上，本文从理论和实践两个层面来解决中古语料分词中出现的问题：首先针对中古汉语中易发生分词不一致的字串，制定并优化分词规范，基于此分词规范校准人工分词语料，尽可能减少人工分词中出现的切分错误和分词不一致情况；然后将整理后的语料作为 CRFs 训练语料，引入字符分类、词典标记两种特征，对两个特征分别设置多个特征模板，并进行对比实验，挑选其中分词效果最好的特征及模板；最后，基于已选定的特征及模板，设置两组分词对比实验，分别验证语料类别和分词一致性对中古汉语 CRFs 分词结果的影响。该实验结果可直接服务于中古汉语语料库的建设。

2 分词说明及实验语料

2.1 分词原则

汉语中词与短语的界限往往难以划分清楚，但这并不妨碍人们对语言的理解。同样，在分词时也不必纠结于语言学领域中词与短语的划界，只要在保证系统适用性、语言单位颗粒度合适的情况下，计算机可以正确理解处理语言单位即可。这一理念在词典收词中亦有所体现，即在保证语义理解正确的前提下，并不强行区分词和短语。因此，绝大部分词典所收录词条的范围不仅仅包括词，也包括一些使用稳固的短语、习用语等。也正是基于此，自然语言处理领域中引入分词单位这一概念，用来指代自然语言处理中使用的、具有确定语义和语法功能的基本单位[4]。中古汉语的词汇单位中存在许多介于词和短语之间的字组，在中古汉语语料库构建过程中，这些字组成为产生分词不一致问题的根源之一，严重影响着语料库的构建质量，理清这些字组的分词边界是提高中古汉语语料库构建质量的关键。

“中古汉语词库”[1](简称为“词库”)是中古汉语语料库项目的重要成果之一，它主要涵盖了以下几本词典所收录的词条及义项：《汉语大词典》[5]、《魏晋南北朝词语例释》[6]、《中古虚词语法例释》[7]、《佛学大辞典》[8]、《佛经词语汇释》[9]。它收录了中古时期出现的绝大部分词语，收录总条目超过 54 万条。基于该“词库”，本文在识别分词界限不清的字组类分词单位时，由于存在词汇化时间、程度模糊难以确定的情况，为了综合照顾到分词单位的统一性、可参照性及语义完整性，总体上遵照以下三个原则：

(1) 从宽原则，在不影响语义理解的情况下，对介于词和短语之间的字组，主张从合不从分。

(2) 词典原则，也即黄居仁所说的分词的信级[10]，规定凡收录在“词库”中的语义单位，一律从合。

(3) 词义透明原则，本文规定，词义不透明包括以下 4 种情况：通过隐喻或转喻方式产生新义；意义发生转指；组成成分的意义有脱落现象；词性发生转变。凡满足词义不透明任一种情况的，一律从合。

基于以上三个原则对人工分词过程进行优化，得到的分词结果作为 CRFs 自动分词模型的训练语料。然而，由于训练语料是多人人工分词的结果，以上三个原则也不能完全避免分词不一致现象[3]的产生。因此，本文对训练语料单独进行整理，尽量消除其中存在的分词不一致现象，并设置对比实验，将分词不一致整理前和整理后的语料分别作为训练语料，验证分词不一致对自动分词结果的影响。

2.2 实验语料说明

中古时期流传下来数量最多的文献种类为史书类和佛经类，除此之外，也有较少量民间文学(如小说)、杂著(如农书)等类别的文献。因此，本文主要选取史书、佛经两类语料作为实验语料，为使实验更具代表性，也加入少量小说类语料。分类抽取中古时期的文献语料作为实验语料，结果如表 1 所示。

表 1 中语料均已初步完成人工分词及标注。按照 2.1 节的分词原则，校对表 1 中语料的分词结果，尽量减少其中存在的分词不一致现象，这样就产生了分词一致性整理前后两批语料，将测试语料校对后的结果作为标准分词结果。对下文三个实验语料，分别说明如下。

表 1　语料情况说明表

语料类别	训练语料			测试语料		
	语料来源	字数	总字数	语料来源	字数	总字数
史书类	后汉书(卷 1、34、74; 卷 2、75、38 未完)	70 344	145 292	北齐书(卷 1-4, 开放测试)	27 189	44 979
	三国志(魏书卷 1-3; 卷 4 未完; 吴书卷 46、卷 49)	62 093		三国志(魏书卷 1-2, 封闭测试)	17 790	
	陈书(卷 1-16; 卷 27-36)	12 855				
佛经类	撰集百缘经	80 588	99 157	百喻经(开放测试)	21 552	35 209
	杂譬喻经二种	18 569		杂譬喻经－失译(封闭测试)	13 657	
小说类	幽明录	36 718	36 718			
总计	281 167			80 188		

(注: “未完”表示实验时, 《后汉书》卷 38 仅有部分完成了人工标注, 但并不影响将已完成部分作为实验语料。下“卷 4 未完”同。)

(1) 在选定特征模板实验环节, 训练语料直接使用表 1 中所有训练语料的人工分词结果, 从 4 本测试语料中各选取一千字, 合起来作为总测试语料;

(2) 在分词一致性对分词结果影响的对比实验中, 分别将分词不一致现象消除前后的语料作为训练语料, 测试语料同表 1;

(3) 在语料混杂度对分词结果影响的对比实验中, 分别将史书、佛经、全部语料作为训练语料, 测试语料同表 1。

3　模型及特征模板选取

基于统计的分词问题可视为文本序列的分类问题, 其核心是在字组中区分出词的起始、中间和结束位置。在给定观察序列的条件下, 分词模型可以统计出整个标注序列的单一联合概率分布[12], 并据此计算或预测出最可能的输出序列。与常用的几个主流模型相比, 如: 隐马尔科夫模型(HMM)、最大熵马尔科夫模型(MEMM)和条件随机场模型(CRFs), 其中 CRFs 具有较为突出的性能, 它不仅克服了 HMM 强独立性假设的局限, 而且解决了 MEMM 标记偏置问题[13]。因此, 本文选择 CRFs 作为实验模型, 在此基础上添加不同特征及特征模板, 选择其中实验效果最好的特征模板, 以提高分词效率。

3.1　特征选取

将 CRFs 模型应用于分词中, 其核心思想是充分挖掘训练语料(输入序列)中构词所用汉字的位置知识[14], 也即特征。特征是 CRFs 分词的核心, 特征选取情况将对分词结果的准确率产生直接影响。但选取的特征并不是越多越好, 选取的特征越多, CRFs 模型提取特征或分词时所需要搜索的数据也越多, 这不仅对机器性能是严峻考验, 也容易使分词结果的准确性受到过度冗余数据的干扰。基于此, 本文选取字符分类和词典标记作为 CRFs 分词特征, 分别说明如下:

(1) 字符分类。字符分类是对语料中字符的粗分类, 这种粗分类无论在现代汉语[15]还是古代汉语[16]中, 都能对分词结果准确率的提高起到较为明显的作用。本文所用字符类别标注集为: T1={ HZ, Punc, SenPunc, CNum, CCNum, D, X }。对应关系为: 汉字(HZ)、普通标点(Punc)、句末标点(SenPunc)、数字(CNum)、干支(CCNum)、“第”字(D)、未识别字符(X)。

由于古代汉语中数字用文字表示, 且不可能完全枚举, 也就不可能完全依靠统计方法实现自动分词, 而数字的表示又有很强的规律性: 多位数字由个位数字组合而成, 个位数字又是封闭的小类, 因此特别对数字、干支(与数字情况相类似)设置了字符类别。而且中古文献使用的为繁体字, 也存在异体字、疑难字等, “未识别字符”统一用来表示程序未能识别出来的文字。字符分类及词典标记采用程序自动标记法。

(2) 词典标记。本文词典标记是动态标记语料中汉字在词典中组词情况的信息。由于 CRFs 分词模型是基于统计学的模型, 它主要依赖词频、上下文构词信息来识别分词单位的界限, 其缺点是难以发现语料中频率较低的词, 且对语料类别有较强依赖性。要克服这些缺点, 有以下两种方法:

①分词时结合规则、词典等信息。由字组词过程中, 可供分词使用的字规则十分有限, 而词典是能用于自动分词的不可忽视的资源, 黄昌宁等[11]主张在语料库标注过程中严格执行“词表驱动”原则, 在没有歧义的情况下, 词表词应

当作为一个完整的切分单位，以保证分词结果的一致性。“统计+词典”的方法不仅能有效利用统计类分词模型便于通过上下文排除歧义、发现新词等优点，也合理运用了已有的语言资源，一定程度上降低了完全基于统计方法需要大规模训练语料的要求，能很好提高自动分词的准确率和效率。

②使用足够多同一类型的训练语料。统计学习方法在分词评测中效果更好[11]，但它需要大量同类型的训练语料，再加上未登录词、组合型歧义的存在，使得统计模型在现代汉语分词结果中的正确率局限于 0.85 左右[17]。而中古语料数量相对较少，语料类型也不统一，再加上分词规范等资源的缺失，采用该种方法，中古汉语 CRFs 分词的正确率只会更低。

基于以上分析，本文在分词时引入词典动态标记语料，将“词库”作为词典标记的依据，引入词典标记标注集为: T2 = { B, M, E, S, W, T, H, F }。标注集中字符含义分别为: 词首字(B)、词中字(M)、词尾字(E)、单字(S)、标点(W)、属于两个词的字(T)、属于三个词的字(H)、属于三个词以上的字(F)。

以《百喻经》首句语言片断“闻如是: 一时佛住王舍城,”为例，动态标记词典信息的具体流程为: 将该语言片断作任意切分，来匹配“词库”收录的词目，除了单字词外，可以匹配到“闻如是、如是、一时、王舍、王舍城”这 5 个词，依据匹配情况可以将语句片断中的汉字分类如下:

(1) “闻、一”两字分别仅出现在“闻如是、一时”的词首位置，因此“闻、一”两字标记为“B”。

(2) “如、是”两字出现在“闻如是、如是”两个词中，“王、舍”两字出现在“王舍、王舍城”两个词中，因此“如、是、王、舍”四字标记为“T”。

(3) “时、城”两字分别出现在“一时、王舍城”的词尾位置，因此“时、城”两字标记为“E”。

(4) “佛、住”两字均在词典中匹配为单字词，故标记为“S”。

此外，标点符号统一标记为“W”。最终，基于“字符分类”、“词典标记”，并将校对后人工分词结果作为标准答案，得到用于 CRFs 实验的语料标准形式如表 2 所示。

表 2 CRFs 语料标记示例

字符	字符类别	词典标记	标准答案
闻	HZ	B	S
如	HZ	T	B
是	HZ	T	E
:	Punc	W	W
一	CNum	B	B
时	HZ	E	E
佛	HZ	S	S
住	HZ	S	S
王	HZ	T	B
舍	HZ	T	M
城	HZ	E	E
,	SenPunc	W	W

3.2 特征模板对比实验

在 CRFs 分词模型中，特征模板是运用特征提取词语边界信息的有效工具。特征模板的设定直接决定能否高效提取到有用的分词信息，对分词结果的好坏也有直接影响。在训练、测试语料中先后加入字符分类、词典标记作为分词特征，并对这两种特征分别选取不同的特征模板，对比不同分词模板作用下训练语料的分词结果，实验结果如表 3 所示。

表 3 加入字符分类、词典标记特征后分词对比

特征 / 统计结果	仅字面信息				字面(1W+2C)+字符分类				字面(1W+2C)+词典				Template-all
	1W	2W	1W+2C	2W+2C	0W	1W	2C	1W+2C	0W	1W	2C	1W+2C	
单字词数	1 710	568	1 918	648	1 300	1 403	1 814	1 384	1 532	1 597	1 866	1 790	1 747
双字词数	970	1 541	866	1 501	1 094	1 042	819	1 045	923	906	803	837	833
多字词数	0	0	0	0	0	0	4	4	16	16	20	20	22
正确分词数	1 127	849	1 223	885	1 120	1 135	1 354	1 147	1 910	1 975	2 033	2 164	2 229
总 P(%)	42.05%	40.26%	**43.93%**	41.18%	46.78%	46.42%	**51.35%**	47.14%	77.30%	78.40%	75.60%	**81.75%**	**85.66%**
总 R(%)	43.02%	32.40%	**46.68%**	33.78%	42.75%	43.32%	**51.68%**	43.78%	72.90%	75.38%	77.60%	**82.60%**	**85.08%**
总 F(%)	42.53%	35.91%	**45.26%**	37.11%	44.67%	44.82%	**51.51%**	45.40%	75.03%	76.86%	76.59%	**82.17%**	**85.37%**
双字词正确数	361	535	334	522	446	414	347	423	592	640	620	634	662
双字词 P(%)	37.22%	34.72%	**38.57%**	34.78%	40.77%	39.73%	**42.37%**	40.48%	64.14%	70.64%	77.21%	**75.75%**	**79.47%**
双字词 R(%)	47.81%	70.86%	**44.24%**	69.14%	59.07%	54.83%	**45.96%**	56.03%	78.41%	84.77%	82.12%	**83.97%**	**87.68%**
双字词 F(%)	41.86%	46.60%	**41.21%**	46.28%	48.24%	46.08%	**44.09%**	47.00%	70.56%	77.06%	79.59%	**79.65%**	**83.38%**

(注: 这里的 W 表示单字, C 为共现字, 数字表示个数, 例如 0W 表示该特征的当前字, 1W+2C 表示相邻字符窗口为±1、二字共现; 列名中 P 表示正确率, R 表示召回率, F 表示 F 值。)

从表 3 可以看出，仅从字面信息提取分词特征时，分词结果中双字词的 F 值与总 F 值之间呈现出负相关关系。当模板为 1W+2C 时分词结果总 F 值最高，此时双字词 F 值虽然最低，但双字词 P 和总 P、双字词 R 和总 R 间比例稳定，其他模板均切分出过多的双字词，同时提高了双字词的错误率和召回率，因此 1W+2C 为字面信息分词效果最好的模板。

字符分类信息能有效地将汉字、标点、数字信息区分开来，因此也就对数字、天干地支等同类别汉字组成的分词单位分词效果尤为明显。加入该特征后，总 F 值提高了约 6%，双字词的 F 值也有所提高，当分词模板为 2C 和 1W+2C 时，还正确切分出了三字及以上的数词。从表 3 可以看出，字符分类信息总体分词效果最好的模板为 2C。

词典是收录词语的权威工具，加入词典标记特征后，分词结果的 F 值得到大幅提升，并且双字词的 F 值与总 F 值间的关系变为正相关，分词效果最好的词典标记模板显然为 1W+2C，在该模板的作用下，分词结果的总 F 值比仅为字面信息时提高了 39.91%。

综上所述，如表 3 中粗体所示，分别选取：字面信息、词典标记特征模板为 1W+2C，也即相邻字符窗口为±1、二字共现；字符分类特征模板为 2C，也即二字共现。选取该模板作为实验的总特征模板，表示如下。

$$\text{Template-all} = (2C)_{\text{字面信息}} + (1W+2C)_{\text{字符分类}} + (1W+2C)_{\text{词典标记}}$$

4 实验设计与评价标准

4.1 实验设计

从理论上来说，CRFs 分词模型中，所选取的实验语料类别越统一，越利于提取到有规律的分词信息。而中古时期的史书、佛经等语料间存在较大语言、词汇方面的差异，语料类别必然也会对 CRFs 分词结果产生一定影响。因此设置以下两组对照实验：

实验 1 分词一致性影响：对人工分词结果进行一致性整理，分别选取分词一致性整理前后的语料作为训练语料，以考察训练语料分词一致性对 CRFs 分词结果的影响。

实验 2 语料混杂度影响：选取分词一致性整理后的语料，分别将史书、佛经单独分词结果与综合起来的分词结果作为训练语料，以考察语料类别混杂度对 CRFs 分词结果的影响。

4.2 评价标准

以校对后的人工分词结果为衡量标准，以准确率、召回率、F 值为评价指标。准确率(Precision，简写为 P)也叫查准率，本文中表示 CRFs 分词结果中的正确率；召回率(Recall，简写为 R)又叫查全率，本文中表示 CRFs 分词结果中正确分词数与标准分词结果的比率；F 值是这两个指标的综合评价。三者的计算公式分别为：

$$P = RW / AW$$

$$R = RW / SW$$

$$F = P \times R \times 2 / (P + R)$$

其中，RW 表示 CRFs 分词结果中正确分词的词目，AW 表示 CRFs 分词结果的总词数，SW 表示人工分词结果中的总词数。

P、R 的取值范围在 0 和 1 之间，F 的取值范围在 P 和 R 之间，这三者数值越接近 1，代表分词效果越好。P 和 R 从不同方面评价了分词模型，F 值则反映了两者的综合评价。

5 数据及分析

在 CRFs 工具中使用上文实验得出的模板 Template-all，对训练语料进行训练，将训练得到的模型按照实验设计分别进行封闭测试和开放测试。

5.1 封闭测试

将训练语料进行分词一致性整理，分别用 CRFs 模型训练整理前后的训练语料，并对测试语料进行封闭测试，以验证分词一致性对 CRFs 自动分词结果的影响，实验结果如表 4 所示。

本实验中，F 值是最重要的性能评价指标，因此在对比实验结果时，主要基于 F 值的变化情况来说明实验结果，如表 4 中粗体字部分，可以得出以下结论：

(1) 对语料进行分词一致性整理后，可以通过上下文、字符分类、词典标记三个特征获取较为准确一致的分词边界信息，其词语的分合规律变得更加有迹可循，因此分词结果的总 F 值得到明显提升，其中史书类语料提升了 15.70%，佛经类语料提升了 12.38%。

(2) 分词单位的字数越多，分词一致性整理对其分词结果的影响越显著。经过一致性整理后，所有词、双字词、多字词，随着词字数的增加，分词结果的 F

表 4　分词一致性对 CRFs 分词结果影响(封闭测试)

训练语料	测试语料	分词结果(CRFs 分词结果的词数与 PRF 值)														
		单字词	双字词	多字词	总 P (%)	总 R (%)	总 F (%)	F 值变化率	双字词 P(%)	双字词 R(%)	双字词 F(%)	F 值变化率	多字词 P(%)	多字词 R(%)	多字词 F(%)	F 值变化率
原语料	史书	7 764	3 263	80	82.05%	85.62%	**83.79%**	↑	81.67%	80.86%	**81.26%**	↑	70.00%	20.59%	**31.82%**	↑
一致后	史书	7 058	3 309	270	99.53%	99.46%	**99.50%**	**15.70%**	99.21%	99.61%	**99.41%**	**18.15%**	98.89%	98.16%	**98.52%**	**66.71%**
原语料	佛经	5 333	2 690	70	88.08%	85.67%	**86.86%**	↑	78.55%	89.95%	**83.87%**	↑	50.00%	26.12%	**34.31%**	↑
一致后	佛经	5 823	2 355	136	99.28%	99.21%	**99.24%**	**12.38%**	99.07%	99.32%	**99.19%**	**15.33%**	91.91%	93.28%	**92.59%**	**58.28%**

值提升效果也越来越明显。这是因为产生分词不一致的均为双字词及多字词，而“词库”以双字词和多字词的收录为主，词典标记强化了双字词和多字词的特征提取。

(3) 多字词的 F 值提升率高达 58%至 67%，远高于总 F 值和双字词 F 值的变化率，这是因为分词单位的字数越多，其在文献中出现的频率就越低，分词结果也就越容易受到分词不一致的干扰。

(4) 分词一致性对佛经类语料影响稍弱于史书类语料。这是由佛经语料的特异性造成的，佛经是翻译过来的文献，语言变异现象较多，也更难提取到规律性的分词边界信息。

选取分词一致性整理后的语料，分别将史书、佛经类语料与综合后的语料作为训练语料，来验证语料类别混杂度对 CRFs 分词结果的影响，实验结果如表 5 所示。

表 5　语料混杂度对 CRFs 分词结果影响(封闭测试)

训练语料	测试语料	分词结果(CRFs 分词结果的词数与 PRF 值)														
		单字词	双字词	多字词	总 P (%)	总 R (%)	总 F (%)	F 值变化率	双字词 P(%)	双字词 R(%)	双字词 F(%)	F 值变化率	多字词 P(%)	多字词 R(%)	多字词 F(%)	F 值变化率
史书	史书	7 764	3 263	80	99.73%	99.71%	**99.72%**	↓	99.61%	99.79%	**99.70%**	↓	99.26%	98.90%	**99.08%**	↓
综合	史书	7 058	3 309	270	99.53%	99.46%	**99.50%**	**0.22%**	99.21%	99.61%	**99.41%**	**0.29%**	98.89%	98.16%	**98.52%**	**0.56%**
佛经	佛经	5 333	2 690	70	99.44%	99.45%	**99.44%**	↓	99.53%	99.32%	**99.42%**	↓	93.43%	95.52%	**94.46%**	↓
综合	佛经	5 823	2 355	136	99.28%	99.21%	**99.24%**	**0.20%**	99.07%	99.32%	**99.19%**	**0.23%**	91.91%	93.28%	**92.59%**	**1.87%**

表 5 反映出，将不同类别语料混同起来用作训练语料时，分词结果 F 值总体呈下降趋势。具体而言，从表 5 可以看出：区分不同类别的语料(史书、佛经)后，总 F 值均提高了不到 0.3%。说明中古时期不同类别语料间虽然存在差异，这种差异对 CRFs 分词结果也造成一定影响，但总体而言影响并不十分大，远远低于分词不一致对分词结果的影响。

5.2　开放测试

按照表 1 对语料进行开放测试，分别将分词一致性整理前后的人工分词语料作为训练语料，进一步验证开放测试中分词一致性对 CRFs 自动分词结果的影响，实验结果如表 6 所示。

表 6　分词一致性对 CRFs 分词结果影响(开放测试)

训练语料	测试语料	分词结果(CRFs 分词结果的词数与 PRF 值)														
		单字词	双字词	多字词	总 P (%)	总 R (%)	总 F (%)	F 值变化率	双字词 P(%)	双字词 R(%)	双字词 F(%)	F 值变化率	多字词 P(%)	多字词 R(%)	多字词 F(%)	F 值变化率
原语料	史书	10 745	5 520	230	80.24%	85.27%	**82.67%**	↑	83.22%	84.51%	**83.86%**	↑	62.17%	20.11%	**30.39%**	↑
一致后	史书	9 834	5 503	513	88.73%	90.61%	**89.66%**	**6.98%**	89.71%	90.82%	**90.26%**	**6.40%**	71.73%	51.76%	**60.13%**	**29.74%**
原语料	佛经	8 482	4 203	76	92.30%	88.46%	**90.34%**	↑	81.51%	94.72%	**87.62%**	↑	60.53%	52.87%	**56.44%**	↑
一致后	佛经	9 113	3 875	84	95.35%	93.61%	**94.47%**	**4.13%**	89.91%	96.32%	**93.01%**	**5.38%**	65.48%	63.22%	**64.33%**	**7.89%**

表 6 中可得到的与封闭测试一致的结论不再赘言，与之不同的是：对训练语料进行一致性整理后，虽然没有封闭测试提升效果那么明显，但开放测试分词结果的总 F 值仍然得到明显提升，

其中史书类语料提升了 6.98%，佛教类语料提升了 4.13%。

选取分词一致性整理后的语料，分别将史书、佛经类语料与综合后的语料作为训练语料，以进一步验证开放测试中语料类别混杂度对 CRFs 分词结果的影响，实验结果如表 7 所示。

表 7　语料混杂度对 CRFs 分词结果影响(开放测试)

训练语料	测试语料	分词结果(CRFs 分词结果的词数与 PRF 值)														
		单字词	双字词	多字词	总 P (%)	总 R (%)	总 F (%)	F 值变化率	双字词 P(%)	双字词 R(%)	双字词 F(%)	F 值变化率	多字词 P(%)	多字词 R(%)	多字词 F(%)	F 值变化率
史书	史书	9 668	5 562	526	88.61%	89.94%	**89.27%**	↑	88.76%	90.82%	**89.78%**	↑	68.82%	50.91%	**58.53%**	↑
综合		9 834	5 503	513	88.73%	90.61%	**89.66%**	**0.39%**	89.71%	90.82%	**90.26%**	**0.48%**	71.73%	51.76%	**60.13%**	**1.60%**
佛经	佛经	9 085	3 902	76	94.82%	93.02%	**93.91%**	↑	89.06%	96.07%	**92.43%**	↑	65.79%	57.47%	**61.35%**	↑
综合		9 113	3 875	84	95.35%	93.61%	**94.47%**	**0.56%**	89.91%	96.32%	**93.01%**	**0.57%**	65.48%	63.22%	**64.33%**	**2.98%**

结合表 5，从表 7 中可以得出以下两个结论:

(1) 在不区分训练语料类别的情况下，开放测试中分词结果的 F 值有所上升。而封闭测试中，分词结果的 F 值却下降了，封闭实验和开放实验的结果呈相反趋势。

(2) 与封闭测试相比，开放测试中总 F 值的变化较大且呈上升趋势，但无论上升还是下降，变化幅度均不超过 1%，与分词一致性对结果的影响相比，小得多且不稳定。

这两个结论说明，语料类别的差异会影响人们的语感，从而加剧分词不一致，因此对语料类别细分类会提高分词效果；在优化分词标准后，一定程度上消除了分词不一致现象，不同类别的语料间词汇差别并没有想象中那么大，对语料类别进行细分后，由于训练语料数量的减少，分词效果反而会打折扣。

综合以上实验，在 CRFs 分词中，按语料类别细分类对分词结果稍有贡献，但在语料总量有限，特别是各类别语料数量不均衡、数量不够大的情况下，由于细分语料就相当于减少了单个实验的训练语料，反而会降低分词效率。而分词是否一致是评测训练语料质量好坏的重要标准之一，在很大程度上影响着分词结果的整体质量。

6　结　语

本文针对中古这一特殊时期的多种语料，制定分词原则，优化了分词过程，然后通过人工校对，尽量减少语料中存在的分词不一致现象。引入字符分类及词典标记特征，通过 CRFs 对比实验选取分词效果最好的模板，实现词典与统计相结合的自动分词方法。可以看出，字符分类、词典标记特征有效利用了汉字构词、已有的中古词库信息，加入字符分类和词典标记特征后，分词结果的总 F 值分别提高 5%和 35%以上，不仅节省了人力，也获得了更好的分词效果。本文实验证明：中古汉语分词一致性对自动分词结果的影响十分显著，对训练语料进行一致性整理后，封闭测试中分词结果的总 F 值均提高了 10%以上，开放测试中分词结果的总 F 值提高了 3%至 7%；区分不同语料对分词结果的影响较小，不足 1%，但由于中古语料数量有限，细分语料类别必然会造成单个实验训练语料的减少，反而可能会降低分词效率。因此在处理好分词一致性的前提下，中古汉语自动分词不必区分处理不同语料。

本文一致性规范的对象仅限于双字词，多字词的处理效果虽然提升显著，但在开放测试中仅有 60%左右，仍然不十分理想，在未来的工作中笔者将考虑:

(1) 研究多字词的分词不一致情况，并制定相应的分词规范，进一步提高训练语料的分词一致性;

(2) 完善“词库”，增强其对中古语料中分词单位的覆盖面;

(3) 在词典标记特征中加入最长匹配标记，提高低频多字词的识别率。

参考文献:

[1] 化振红. 深加工中古汉语语料库建设的若干问题[J]. 西南大学学报: 社会科学版, 2014, 40(3): 136-142. (Hua Zhenhong. Some Problems in the Deep Processing of the Medieval Chinese Corpus Construction[J]. Journal of Southwest University: Social Science Edition, 2014, 40(3): 136-142.)

[2] 王嘉灵. 以《汉书》为例的中古汉语自动分词[D]. 南京: 南京师范大学, 2014. (Wang Jialing. The Medieval Chinese Automatic Segmentation Using the “Han Shu” as an Example [D]. Nanjing: Nanjing Normal University, 2014.)

[3] 王晓玉, 董志翘. 中古汉语分词不一致原因探讨[J]. 汉语史研究集刊, 2015, 19: 20-33 (Wang Xiaoyu, Dong Zhiqiao. The Investigation of Middle Ancient Chinese Word Segmentation’s Inconsistency[J]. The Collected Papers of the Chinese History Study, 2015, 19:20-33.)

[4] GB-T13715-1992. 信息处理用现代汉语分词规范[S].北京: 中国标准出版社, 1993. (GB-T13715-1992. Contemporary Chinese Language Word Segmentation Specification for Information Processing [S]. Beijing: China Standard Press, 1993.)

[5] 罗竹风,等. 汉语大词典[M]. 上海: 上海辞书出版社, 2011. (Luo Zhufeng, et al. The Great Chinese Dictionary [M]. Shanghai: Shanghai Lexicographical Publishing House, 2011.)

[6] 蔡镜浩. 魏晋南北朝词语例释[M]. 南京: 江苏古籍出版社, 1990. (Cai Jinghao. Wei, Jin, Southern and Northern Dynasties Words and Expressions [M]. Nanjing: Jiangsu Ancient Books Publishing House, 1990.)

[7] 董志翘, 蔡镜浩. 中古虚词语法例释[M]. 长春: 吉林教育出版社, 1994. (Dong Zhiqiao, Cai Jinghao. Middle Ancient Function Words and Expressions [M]. Changchun: Jilin Education Publishing House, 1994.)

[8] 丁福保. 佛学大辞典[M]. 北京: 中国书店出版社, 2011. (Ding Fubao. Buddhist Dictionary [M]. Beijing: China Bookstore Publishing House, 2011.)

[9] 李维琦, 蒋冀骋. 佛经词语汇释[M]. 长沙: 湖南师大出版社, 2004. (Li Weiqi, Jiang Jicheng. Sutras Words Explanations [M]. Changsha: Hunan Normal University Publishing House, 2004.)

[10] 黄居仁, 陈克健, 陈凤仪,等. 《资讯处理用中文分词规范》设计理念及规范内容[J]. 语言文字应用, 1997(1):94-102. (Huang Juren, Chen Kejian, Chen Fengyi, et al. A Segmentation Standard for Chinese Information Processing: Design Criteria and Content [J]. Journal of Applied Linguistics, 1997(1): 94-102.)

[11] 黄昌宁, 赵海. 中文分词十年回顾[J]. 中文信息学报, 2007, 21(3): 8-19. (Huang Changning, Zhao Hai. Chinese Word Segmentation: A Decade Review [J]. Journal of Chinese Information Processing, 2007, 21(3): 8-19.)

[12] 吴琼, 黄德根. 基于条件随机场与时间词库的中文时间表达式识别[J]. 中文信息学报, 2014, 28(6): 169-174. (Wu Qiong, Huang Degen. Temporal Information Extraction Based on CRF and Time Thesaurus [J]. Journal of Chinese Information Processing, 2014, 28(6): 169-174.)

[13] 段宇锋, 朱雯晶, 陈巧, 等. 条件随机场与领域本体元素集相结合的未登录词识别研究[J]. 现代图书情报技术, 2015(4): 41-49. (Duan Yufeng, Zhu Wenjing, Chen Qiao, et al. The Study on Out-of-Vocabulary Identification on a Model Based on the Combination of CRFs and Domain Ontology Elements Set[J]. New Technology of Library and Information Service, 2015(4): 41-49.)

[14] 修驰. 适应于不同领域的中文分词方法研究与实现[D]. 北京: 北京工业大学, 2013. (Xiu Chi. The Research and Implementation of Chinese Word Segmentation for Different Domains [D]. Beijing: Beijing University of Technology, 2013.)

[15] 宋彦, 蔡东风, 张桂平, 等. 一种基于字词联合解码的中文分词方法[J]. 软件学报, 2009, 20(9): 2366-2375. (Song Yan, Cai Dongfeng, Zhang Guiping, et al. Approach to Chinese Word Segmentation Based on Character-Word Joint Decoding[J]. Journal of Software, 2009, 20(9): 2366-2375.)

[16] 石民, 李斌, 陈小荷. 基于 CRF 的先秦汉语分词标注一体化研究[J]. 中文信息学报, 2010, 24(2): 39-45. (Shi Min, Li Bin, Chen Xiaohe. CRF Based Research on a Unified Approach to Word Segmentation and POS Tagging for Pre-Qin Chinese[J]. Journal of Chinese Information Processing, 2010, 24(2): 39-45.)

[17] Zhao H, Kit C Y. An Empirical Comparison of Goodness Measures for Unsupervised Chinese Word Segmentation with a Unified Framework[C]//Proceedings of IJCNLP 2008, Hyderabad, India. 2008: 9-16.

作者贡献声明:

李斌: 提出研究思路, 软件编程实现, 修改论文质量;

王晓玉: 负责进行实验, 获取并分析数据, 论文起草和修改。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。

[1] 王晓玉, 李斌. 结合词典的中古汉语 CRFs 分词测试数据.

收稿日期: 2017-03-14
收修改稿日期: 2017-04-20

Automatically Segmenting Middle Ancient Chinese Words with CRFs

Wang Xiaoyu　Li Bin
(School of Chinese Language and Literature, Nanjing Normal University, Nanjing 210097, China)

Abstract: [**Objective**] The purpose of this paper is to explore the influence of the word segmentation consistency and the corpus types in Middle Ancient Chinese (MAC). It tries to improve the accuracy and efficiency of the automatic word segmentation, a basic procedure in processing ancient Chinese, based on the CRFs model. [**Methods**] First, we optimized the segmentation principles for MAC historical records, Buddhist scriptures and novels. Then, we combined the CRFs model with dictionary to reduce the segmentation inconsistency in the manual procedures. Finally, we added two features to the CRFs model (i.e. character classification and dictionary information), and identified the best word segmentation template by comparison experiments. [**Results**] The F-score was higher than 99% in the closed test, while it was from 89% to 95% in the open test. [**Limitations**] The segmentation consistency was improved on the words with two characters, and more studies were needed on the segmentation of words with more than three characters. [**Conclusions**] The proposed method could effectively improve the accuracy of automatic word segmentation for mediaeval Chinese corpus.

Keywords: Conditional Random Fields Model　Segmentation Consistency　Middle Ancient Chinese Word Segmentation

IMLS 资助 Educopia 和 UNC SILS 研究以改进原生数字资源存档工作流的开放源码软件

Educopia(一个非营利性的组织, 旨在促进文化、科学和学术机构之间的合作)和北卡罗莱大学信息和图书馆科学学院(UNC SILS)已从美国博物馆和图书馆服务研究所(Institute of Museum and Library Services, IMLS)获得价值超过 68 万美元的资助, 用于支持 OSSArcFlow 项目, 该项目旨在调研图书馆、档案馆和博物馆采用开源工具的情况并提供支持。研究团队将与 12 家合作机构进行合作, 共同研究、设计和测试实施三大领先的开源软件(OSS)技术: BitCurator 环境, ArchivesSpace 和 Archivematica。

通过与多种规模和类型的机构合作, 该项目的调查人员将能够收集到重要的意见, 这将使许多图书馆和档案馆受益。最终, 所有项目成果包括陈述、工作流程、总结结果、培训模块和指南, 都将得到广泛传播, 以帮助其他机构成功地采用开源的数字化策略和保存工具。

北卡罗莱大学信息和图书馆科学学院教授, 项目主要研究人员 Christopher Lee 说: “我们的目标是使得图书馆、档案馆和博物馆实施数字化管理工具的艰巨任务变得更加容易。该项目将发现并支持更有效率、更有效的数字化保存工具, 来促进图书馆和档案馆的努力, 从而确保全人类对日益增长的原生数字化文化遗产的持续访问。”

该项目的合作机构包括杜克大学、麻省理工学院、纽约公立图书馆、纽约大学、莱斯大学和斯坦福大学等。

(编译自: https://sils.unc.edu/news/2017/OSSArcFlow)

(本刊讯)